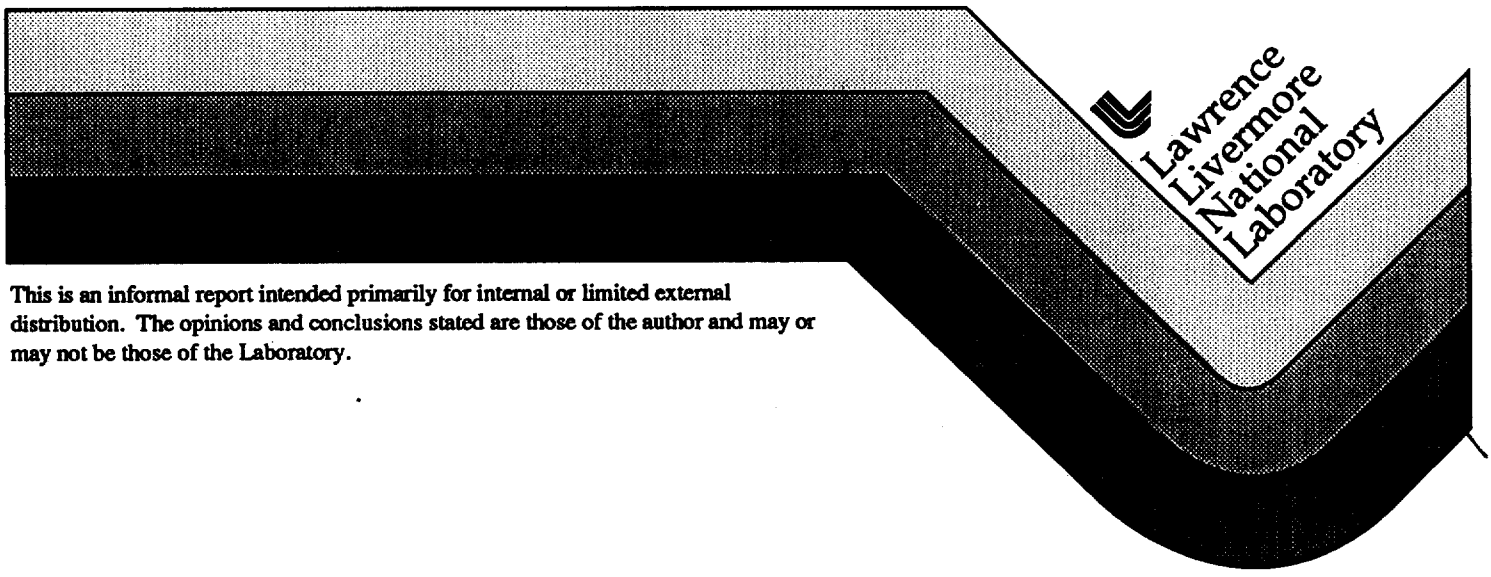
UCRL-ID-120644 Rev 1

PDDP, A Data Parallel Programming Model

Karen Warren

June 1995



PDDP, A Data Parallel Programming Model*

Karen H. Warren
Lawrence Livermore National Laboratory
Livermore, California 94551

Abstract: PDDP, the Parallel Data Distribution Preprocessor, is a data parallel programming model for distributed memory parallel computers. PDDP implements High Performance Fortran compatible data distribution directives and parallelism expressed by the use of Fortran 90 array syntax, the FORALL statement, and the WHERE construct. Distributed data objects belong to a global name space; other data objects are treated as local and replicated on each processor. PDDP allows the user to program in a shared-memory style and generates codes that are portable to a variety of parallel machines. For interprocessor communication, PDDP uses the fastest communication primitives on each platform.

1 Introduction

Parallel programming languages have traditionally been complex and architecture dependent. All but the simplest message passing system preclude portability. Message passing itself has been described as the assembly language of parallel computers.

In 1992, members of the Massively Parallel Computing Initiative project at Lawrence Livermore National Laboratory (LLNL) proposed writing an experimental translator that would allow the user to code in a high-level Fortran-based SPMD language. The resulting code would make efficient use of MIMD computers with non-uniformly accessible memories. The project goals were to examine the technology involved and to investigate the merits of such a language, including whether such an architecture-independent language could indeed be used efficiently on any parallel computer with distributed memory. A valuable additional benefit for both implementors and users would be to gain experience in parallel processing with a high-level programming model.

In this paper, we present the resulting language model, PDDP, the Parallel Data Distribution Preprocessor. We present the syntax and semantics of PDDP, describe its implementation, discuss portability issues, and present data on its performance.

*Work performed under the auspices of the U. S. Department of Energy by the Lawrence Livermore National Laboratory under contract No. W-7405-Eng-48.

2 Background

PDDP is a hybrid of PFP [1], a Parallel Fortran Preprocessor used at LLNL, and Fortran D [2], a research compiler from Rice University. Fortran D provides an extensive set of declarations for distributing data across processor memories and also serves as a base for the HPF [3] distribution directives. Over the past two years, the High Performance Fortran Forum has focused on the need for a high-level Fortran parallel programming model. The resulting HPF language specification is a published model ready for implementation [3]. Because PDDP contains a subset of HPF, PDDP codes are easily converted to HPF.

Its other predecessor, PFP, is a task-oriented parallel Fortran programming language. In the PFP programming model, all of the processors, requested at run-time and referred to as a *team*, enter the main routine in parallel. The user directs this team through the application with the option of dividing the team into subteams to perform tasks in parallel. PFP offers the familiar shared-memory programming model elements, including barriers and *shared* and *private* storage attributes for variables. In a similar manner, all of the processors requested at run-time execute each statement of a PDDP code except for master blocks and parallel code segments. The processors execute the code statements, in a semi-synchronous manner, uninhibited by implicit synchronization in any of the constructs. PFP provides an explicit synchronization tool, a `barrier` statement. Currently, PDDP does not implement team splitting for parallel tasks, rather parallelism is expressed in the HPF FORALL, the Fortran 90 array syntax, and WHERE statments.

3 PDDP Syntax

PDDP consists of a one pass parser-translator and a run-time library. The parser accepts a superset of FORTRAN 77 statements. For each source statement, the parser builds a parse tree used to generate FORTRAN 77 code. User declarations include a subset of HPF **DISTRIBUTE**, **TEMPLATE**, and **ALIGN** specification directives. The parser builds a symbol table of declared scalars, arrays, distributions, common blocks, and subroutines. For array and distribution declarations, it records the number of dimensions and extents. It recognizes array-slice and whole-array syntax as well as individual distributed array accesses. It recognizes distributed arrays used in subroutine arguments and common statements. The use of Fortran 90 [4] array syntax, the **WHERE** construct, and the HPF **FORALL** statement imply parallel execution by the members of the team.

The PDDP parser recognizes the following HPF distribution specification directives:

```
TEMPLATE
DISTRIBUTE
ALIGN
```

Together they indicate the mapping of the data to the processor memories. An abstract array is first declared using the **TEMPLATE** statement. It is partitioned among the processors using the **DISTRIBUTE** statement along with a HPF data distribution type for each dimension:

- **BLOCK** places successive array elements on the same processor, moving to the next processor when the block size, equal to the extent divided by the number of processors, has been used up.
- **CYCLIC** causes successive elements of the array to be placed on successive processors in the system, wrapping around after the last processor.
- The degenerate distribution "*" leaves the entire dimension on a single processor.

Actual arrays are associated with the abstract template using the **ALIGN** statement.

Distributed arrays are globally accessible and are distributed across the processor memory regions. For communication purposes, PDDP also provides global objects that are not distributed but may be accessed by all processors. These are referred to as "shared-only" objects; their names do not occur in **ALIGN** statements. There are two PDDP storage class modifiers:

```
shared
private
```

The **shared** modifier must be used in all declarations of distributed and shared-only objects. By default, non-distributed objects, or those declared using the attribute, **private**, are replicated in local memories.

To indicate execution by only one processor, the user places statements within the following construct:

```
MASTER
ENDMASTER
```

To synchronize the team of processors, the

```
BARRIER
```

statement is available.

PDDP recognizes the Fortran 90 **WHERE** construct with an optional **ELSEWHERE**:

WHERE *logical-array-expression*
    assignment statement

ELSEWHERE
    assignment statement

ENDWHERE

The PDDP **FORALL** statement is similar to the HPF **FORALL**. It takes the form

FORALL (*index-specifications[,*
    scalar-mask-expression])
    assignment-statement

where *index-specifications* takes the form:

index-name = subscript : subscript [: stride]

FORALL may be used for a scatter operation if the expression used to designate the resulting location is entirely local. For example, in the following statement, `index(i)`, must be a local array:

```
FORALL (i = 1:100) (x(index(i)) = i**2
```

PDDP recognizes the Fortran 90 syntax for the global reduction functions **max**, **min**, **sum**, **product**, **any**, **all**. For each, it generates inline code that causes each processor to calculate its local result and store it in a shared-only array. After a barrier, processor 0 calculates the result as a global scalar. Each processor then makes a local copy. PDDP also recognizes the **cshift** function. Inline coding restricts the user to one reduction per statement.

A PDDP code converts easily to HPF. The programmer should preface **master**, **endmaster**, and **barrier** statements with the column 1 tag "**CPDDP$**" so that HPF will ignore them. The **shared** and **private** attributes used in declarations can easily be removed with the use of macros. Then HPF will compile the PDDP source code.

4 PDDP Semantics

Generated code consists of FORTRAN 77 statements. PDDP translates each distribution and **TEMPLATE** declaration into a call to a library routine that assigns to the distribution an ID tag and writes a local table of the necessary information. Distributed arrays must be dynamically allocated on the local heap. Shared-only arrays used as subroutine arguments or in common statements must also be allocated. For each distributed array, PDDP translates the appropriate **ALLOCATE** statement executed by each processor into calls to library routines. These routines give the array an ID tag, allocate the appropriate amount of local memory, and build a local database linking the array to its distribution information and to an address map. The address map gives the starting address of the memory allocated in each processor's memory. The processors use these addresses for requesting remote data (see Table 1).

Table 1. Array Database

Array Data
ARRAY TAG
DISTRIBUTION TAG
number bytes/element
rank
global bounds
local bounds
ADDRESS MAP PTR

Distribution
DISTRIB TAG
rank
extent/dim
distrib type/dim
no. procs/dim

ADDRESS MAP	
Proc	Address
0	0x000200
1	0x000120
2	0x000040
3	0x000220
4	0x000120
...	...

PDDP codes use the "owner-computes" rule for parallel execution of assignment statements: the owner of the left-hand side element executes the assignment for that element. PDDP initially assumes that the right hand side is remote; however, it will not issue a **get** on distributed-memory machines if the processor number of the requested address is the same as the requesting processor. Generated declarations include the pointers and variables needed by PDDP to express a Fortran 90 array statement as do loops whose bounds are the indices for the local portion of the left-hand side array. There is no restriction to the number of seven possible dimensions that can be distributed or the extent of any distributed dimension. For multidimensional left-hand side arrays, the do loops are nested with the order going from the left to the right-most dimension.

Because the left-hand side owner is determined at run-time, PDDP allows dynamic array sizes and varying numbers of processors. For a left-hand side scalar reference to a distributed object, PDDP simply inserts a call to routines that determine the owning processor. Only the owning processor executes the statement. Note that this is substantially different from a scalar reference to a non-distributed data item. In the first case, the statement is executed via owner-computes. In the latter case, all team members execute the statement. The user must be careful with statements that contain data dependency between left- and right-hand sides.

Subroutine linkage in PDDP ensures consistency across subroutine boundaries. With the exception of local routines (see Section 5), array slices are not allowed as arguments in subroutine calls. To pass entire distributed arrays to other modules, PDDP recognizes the use of whole array syntax used in subroutine calls or in common statements. The called subroutine must align a distributed argument to a template with the same distribution as specified in the calling routine. Automatic redistribution on subroutine entry is not supported. Rather than sending a valid address as the argument to a routine, PDDP actually passes the ID tag associated with the array. (The tag is created in the allocation process.) Similarly, it is the ID tag that is actually used in a common block. In the receiving routine, the ID tag allows the module to access information on the data object by using the run-time support routines (see Section 5). The tag is selected so as to cause a fault if referenced without proper declaration and query of the run-time routines. This helps to reduce the number of errors that can be made by new users.

4.1 Optimizations

The PDDP parser recognizes matching array syntax and distribution for left- and right-hand expressions and avoids the time-consuming calculation of the owner. It also avoids divisions involving a stride of 1. If the rank of the left-hand and right hand arrays are unequal and the extra dimensions have a degenerate distribution, the parser also omits generating code that performs calculation of the owner.

Because PDDP is a source-to-source language translator, it is limited in the range of possible optimiza-

tions. It is dependent on the backend compiler optimizer for many performance improvements.

5 Run-time Library

As Nitzberg and Lo [5] point out, a useful distributed shared-memory system must automatically transform shared-memory access into interprocess communication. To achieve this, it is necessary for each processor to have knowledge of the mappings of the distributed arrays so that non-local memory may be accessed and the owner of array elements may be determined. As mentioned above, the PDDP parser generates calls to the run-time library routines that build and access linked tables that make up a local database. The number of processors is a run-time parameter. The data is used to determine the run-time owner, the bounds for the generated do loops, and the location of each right-hand-side distributed object in terms of processor number and offset from the starting address on that processor

Given the global iteration set specified by the user in array syntax and the knowledge of the resident elements from the database, PDDP uses Euclid's extended algorithm [6] to calculate the intersection, a set of local loop indices for a processor. For block distribution, the run-time module takes shortcuts in calculation of the owner. The local array address map allows PDDP to express the actual assignment statement in terms of pointers and offsets, and optional processor numbers for the right hand side.

To demonstrate the use of the database and run-time libraries, consider the following PDDP code. Different distribution are used on left- and right-hand sides to demonstrate the use of the library:

```
integer nx
real x(nx), y(nx)
template t1(nx)
distribute t1(block)
align x with t1
template t2(nx)
distribute t2(cyclic)
align y with t2

x = y
```

Below is the PDDP generated pseudo code:

```
  pointer (ptr0, local_mem)
  pointer (ptr_rh, remote_mem)
  integer address_map_x(no_procs)
  integer address_map_y(no_procs)
     ...
c    loop setup:
     ptr0 = address_map_X(myproc)
     lo_indx = 1
     hi_indx = nx
     stride  = 1
     call get_local_indx(X_id,
    >              lo_indx, hi_indx, stride)
c    lo_indx, hi_indx, stride are now local bounds
     stride_rh  = stride
     lo_indx_rh = lo_indx
     do indx1 = lo_indx, hi_indx, stride
        offset     = mod(indx1, no_procs)
        proc_no_rh = mod(lo_indx_rh, no_procs)
        offset_rh  =  div(lo_indx_rh, no_procs)
        ptr_rh     =  address_map_Y(proc_no_rh)
c     assignment statement:
        local_mem(offset) =
    >     get(proc_no_rh, remote_mem(off_set_rh))
        lo_indx_rh = lo_indx_rh + stride_rh
     enddo
```

In addition to supplying routines that are called by the generated code to calculate the owner, the run-time library supplies routines for the user and debugging tool to query the database. Inquiry functions give the rank and global and local bounds of a distributed array as well as the size in terms of the number of elements of the local block of memory, and the starting address of the local block of memory.

One of the library routines gives the starting address and size of the local array block and thus allows the user to pass the local array section to local routines. Other routines supply the processor number and total number of processors.

6 I/O

PDDP does not offer parallel input/output. Write and read statements must be placed within `master`, `endmaster` blocks. and the variables used must either be local or shared-only (i.e., not distributed). This is obviously awkward and a definite weakness in most high-level parallel programming languages.

7 User Interface

PDDP accepts files with the suffix `.pddp`, as well as `.PDDP`, `.F`, `.f` and `.o`. PDDP passes options other than those directed to the parser on to the compiler and loader [7]. For example:

```
pddp -o code.x -g obj.o code.pddp
```

In the above example, PDDP translates the file `code.pddp` into `code.f`, which is passed to the FORTRAN 77 compiler along with the option `-g`. Then PDDP passes the resulting `code.o` along with `obj.o` to the loader. The option `-barrier` may be used to place a barrier after each array syntax statement translation to test for race conditions. This puts PDDP into a SIMD-like mode for array operations only.

Use of the `-nodist` option causes PDDP to ignore data distributions statements, substituting shared-memory declarations. The resulting code is a shared-memory program that can be used for timing and debugging.

Debuggers can display the generated FORTRAN 77 code or, in the case that the native compiler recognizes lines beginning with "`#[line]`", the debugger can display the original user code. This was advantageous for PDDP users on the BBN TC2000. They were able to use the Totalview X-window debugger to easily debug their PDDP codes. In either case, the run-time library provides debugging functions to display the values of a distributed array, array slice, or designated array element. Indices, bounds, and resident processor may also be printed. To see the memory configuration of a given distributed array, `pddp_config` displays the processor number, local lower and upper bounds, stride, and distribution type of the entire array.

8 Portability

One of the most important characteristics of PDDP is its portability. It is designed to generate code for any parallel computer with shared memory or distributed memory that has the capability, either in hardware or software, for one processor to request and receive data located in another processor's memory.

When porting PDDP to the various platforms, we had to consider several issues besides the major one of internodal communication. These included the peculiarities of the native FORTRAN 77 compiler. For example, cf77 does not allow "`#[line]`" line directives.

For shared-memory machines, we had to decide how to implement distributed memory, and on those machines with only distributed memory we had to decide how to implement shared-only memory.

Because all of the processors execute the entire code, we had to arrange for all of the processors to be forked and ready to execute the first statement.

8.1 Platforms

On architectures with hardware support for remote memory references, such as the BBN TC2000 and the CRI T3D, the task of writing a compiler for the data parallel programming model is greatly simplified. With the owner-computes rule in effect, the processor that handles the computation for a section of an array receives the remote data that it needs through the use of remote memory reference support. The nature of the compiler is that of a finite state engine that handles all of the actions for the processor that is performing the work. To perform efficiently on other architectures, PDDP uses the fastest available means of communication to obtain remote data.

PDDP was initially developed on the BBN TC2000, a computer with distributed but globally addressable memory. PDDP currently is available on the CRI T3D, the Meiko CS-2, and the SGI Power Challenge.

Each BBN processor had a 12 MB low-latency "local" memory and thus resembled a distributed-memory architecture. Each processor also contributed 4 MB to an interleaved shared-memory wherein successive cache lines were placed on successive processors and wrapped around. Because there was a single address space, it also resembled shared-memory. The hardware handled non-local accesses, so there was no need for explicit message passing. On the BBN, a run-time library module called "niam" started first; this routine forked the necessary processors and then called the user's main program. When the main program returned, "niam" terminated the other processors and then itself exited.

On the T3D and Meiko CS-2, the system takes care of starting up all of the requested processors. On these two platforms, processor 0 serves as the resident of shared-only objects. On the Meiko this is much less efficient than the interleaved shared-memory on the BBN.

Each node on the Meiko has 128 MB of memory. The Meiko has a 70 MHz multistage fat tree interconnect, an Elite network switch, and an Elan communications processor. The Elite Switch is an eight-way crossbar switch allowing input/output pairing without contention. Usable bandwidth is 50 MB/s/link in each direction. To read remote data, PDDP uses `fetch` from the Elan Widget Library. The Elan Widget communications library views the address spaces of processors as distributed global memory and explicitly addresses non-local memory by network DMA operations.

Memory on the CRI T3D is globally accessible and physically distributed, 64 MB per processor. Remote

memory referencing is done with a replicated virtual memory address space and separate tracking of processor indices. The 128 processors of the T3D are linked with a 3D torus communications network capable of low-latency data transfers of over 140 MB/sec node to node. Peak per processor performance is 150 Mflops. In a manner similar to PDDP, the CRI data parallel programming model, CRAFT [8], allows the user to view the distributed memory as logically shared and sets the default storage type to private. However, CRAFT restricts the user to powers of two in the distributed dimensions. On the T3D, PDDP allocates memory on the shared-memory heap and uses **shmem_get** from the SHMEM library to access right hand side data. **shmem_get** does a blocking transfer of data from the remote address into the local address using remote loads. It would be advantageous to do a put instead, but that is not compatible with the owner-computes rule. To avoid segmentation violation errors when accessing remote addresses on the T3D, we allocate the same amount of memory on each processor for a distributed array regardless of whether it is used.

Although the PDDP model is directed to non-uniform access distributed-memory architectures, PDDP can also be used on computers with a single shared memory. PDDP was ported to the SGI computer to provide a developmental platform for massively parallel computer users. On the SGI, PDDP forks the desired number of processors, which executes the code as a team. It ignores the shared and private attributes and translates the use of Fortran 90 syntax, `FORALL`, and `WHERE` statements, into do loops in which the indices are interleaved among the processors in a wrap around manner.

8.2 Performance

To demonstrate the performance of PDDP, we present results from four codes in our benchmarking suite (see Tables 2, 3, 4, and 5). We include times from the CRI CRAFT model and the Portland Group HPF, version 1.1-1. The Gaussian non-pivoting elimination solver uses a `CYCLIC` distribution for the second dimension of the matrix. The NAS benchmark implicit PDE solver, LU, for 5 coupled, non-linear partial differential equations uses a `BLOCK` distribution in the last two dimensions. The highly parallel shallow water code is a two dimensional finite difference algorithm on a 512×512 grid. The second dimension is distributed in a blockwise manner across the processors. The data in the quantum lattice gauge code are four and six dimensional arrays of complex variables representing 3×3 arrays in four-dimensional space. The arrays are distributed `BLOCK, BLOCK, BLOCK` in the three rightmost dimensions. A large portion of the calculation is the multiplication of 3×3 matrices.

In the following tables, "N" indicates the number of processors.

Table 2. Gaussian Elimination Algorithm (non-pivoting) 1024×1024

	Time (sec)				
N	T3D PDDP	T3D CRAFT	T3D PVM	MEIKO PDDP	MEIKO PGHPF
16	48	22	17	85	678
32	32	17	12	173	1160
64	26	16	12	353	2780
128	23	19	12		

Table 3. Shallow Water (512×512) 50 iterations

	Time (sec)			
N	T3D PDDP	T3D CRAFT	MEIKO PDDP	MEIKO PGHPF
16	30.0	9.3	82	13
32	16.7	4.8	64	11
64	10.2	2.5	54	15
128	7.0	1.4	47	

Table 4. LU (64×64)

	Time (sec)	
N	T3D PDDP	MEIKO PDDP
16	4480	9285
32	2367	5811
64	1235	3635
128	773	2646
256	421	

Table 5. Quantum Lattice Gauge Code
1 Loop for 2048 elements

	Time (sec)	
N	T3D PDDP	MEIKO PDDP
2	150.6	194.0
4	79.5	139.0
8	43.6	114.8
16	24.3	89.8
32	21.	130.
64	11.	123.

On platforms that do not efficiently support remote memory referencing, e.g., the Meiko, latency of short messages can be a limiting factor. The read bandwidth on the T3D is 2 ns versus 30 ns on the Meiko.

On a platform such as the Meiko, if one cannot repackage communications into long messages and transmit them prior to need, performance suffers. To date, this prefetch has been a task treated by hand in programs using the message passing programming model. In the case of high-level languages, it would be advantageous to accomplish this transparently under control of the compiler. William Carlson from SRC has recently developed an AC compiler [9] for the T3D that does a prefetch and shows good results.

Under the control of PDDP, processors act as a vector unit for the duration of the loop and consequently would greatly benefit from having the performance characteristics of a conventional vector processor.

9 Conclusions

It is evident from our numbers that some form of hardware support for accessing remote memory is necessary for good performance from a high-level parallel programming language. If this support is not present, then some form of a prefetch mechanism is necessary. PDDP is unique in its utilization of hardware support for accessing remote addresses. Implementation of a shared memory programming style itself has proven to be a fundamental feature of massively parallel programming environments. Vendors are striving to place this functionality in the hardware itself.

In evaluating a model such as PDDP, we need to consider the effort required to write a code in a language such as PDDP and compare it to that of porting a code written in message passing, analyzing its performance on the target architecture, and tuning it in some cases via assembly language to obtain reasonable performance. These latter tasks take considerable time and effort and require in-depth knowledge of the target architecture.

A reasonable fraction of this performance can be achieved by using a high-level programming model such as PDDP. While the code does not perform as well as vendor specialized software, scientists prefer the portability trade-off gained. PDDP users can attain reasonable performance with considerably less work than is required today on massively parallel systems. In addition, portability gives application programmers the benefit of single-source maintenance.

PDDP is a research vehicle and a simple language. Nevertheless, we have shown that it is possible to program codes for parallel computers in a high-level language, avoiding the complexities of message passing and achieving satisfactory performance with one source code on multiple parallel platforms.

10 Acknowledgements

Other contributors to the PDDP project have been Brent Gorda, Andrew Ingalls, James Stichnoth, Alan Riddle, Bor Chan, and Paul Lu.

References

[1] Karen H. Warren, Brent Gorda, and Eugene D. Brooks III, *Programming in PFP*, Lawrence Livermore National Laboratory, Livermore, CA, UCRL-MA-107028, 1991.

[2] G. Fox, S. Hiranandani, K. Kennedy, C Koelbel, U. Kremer, C. Tseng, and M. Wu, *Fortran D Language Specification*, Dept. of Computer Science, Rice University, Technical Report TR90-141, December, 1990.

[3] High Performance Fortran Forum, High Performance Fortran Language Specification, Rice University, Houston, Texas, Version 1.1, November 10, 1994.

[4] ISO. *Fortran 90*, May 1991. [ISO/IEC 1539: 1991 (E)].

[5] Bill Nitzberg and Virginia Lo, "Distributed Shared Memory: A Survey of Issues and Algorithm", *Computer*, August 1991, pp. 52-60.

[6] Donald E. Knuth, *The Art of Computer Programming*, Volume 1/Fundamental Algorithms, Addison-Wesley, 1973.

[7] Karen Warren, "PDDP: A Parallel Data Distribution Preprocessor", Lawrence Livermore National Laboratory, Livermore, CA, pp. 42-51 in MPCI Yearly Report 1992: Harnessing the Killer Micro, UCRL-ID-107022-1992.

[8] Cray MPP Fortran Reference Manual, Cray Research, Inc. SR-2504 6.1, 1994.

[9] William Carlson and Jesse Draper, "Distributed Data Access in AC," Bowie, MD, IDA Supercomputing Research Center, December 14, 1994.